Journal of
Economic and Social Thought
www.kspjournals.org
Volume 3 September 2016 Issue 3

An Introduction to Alternative Methods in Program Impact Evaluation

By Cuong NGUYEN †

Abstract. During the recent years, researchers as well as policy makers have been increasingly interested in impact evaluation of development programs. A large number of impact evaluations have been developed and applied to measure the impact of programs. Different impact evaluation methods rely on different identification assumptions. This paper presents an overview of several widely-used methods in program impact evaluation. In addition to a randomization-based method, these methods are categorized into: (i) methods assuming "selection on observable" and (ii) methods assuming "selection on unobservable". The paper discusses each method under identification assumptions and estimation strategy. Identification assumptions are presented in a unified framework of counterfactual and two-equation model.
Keywords: Program impact evaluation, treatment effect, counterfactual, potential outcomes
JEL classification:C40, H43, J68.

1. Introduction

There is a growing interest in impact evaluation of development programs and policies for not only academic researchers but also policy makers. Impact evaluation of a program provides very helpful information for decisions as to whether the program should be terminated or expanded. If a program has no impacts on its participants, it needs to be stopped or revised.

There are several definitions of impact evaluations (White 2006; 2009). The main objective of impact evaluation of a program is to assess whether the program has achieved its objectives of improving outcomes of targeted groups. According to White (2006), most programs have a log frame indicating the program path from inputs to outputs, outcomes and impacts of the programs, and 'any evaluation that refers to impact indicators is thus, by definition, an impact evaluation'. Program impact evaluation methods consist of both quantitative and qualitative methods. In this paper, we will focus discussion on quantitative methods, which are used to measure the impact of a program. The impact of a program on beneficiaries is defined as the change in outcomes of a beneficiary population that can be attributed only to the program.

Unlike experimental studies in medical or physical science, participants are not randomly selected in most socio-economic programs or projects. Simple comparison of outcomes between participants and non-participants in a non-randomized program cannot provide unbiased estimates of the program impact.

† National Economics University, Hanoi, Vietnam; Mekong Development Research Institute, Hanoi, Vietnam; IPAG Business School, Paris, France.
. (84) 904 159 258
. c_nguyenviet@yahoo.com

The difficulty in impact evaluation is also referred as a missing data problem. As mentioned, the impact of a program on an outcome of a participant is defined as the difference between its outcome with the program and its outcome without the program. However, for participants of the program, we can observe only their outcome in the program state, but not their outcome if they had not participated in the program – their counterfactual. Similarly, for non-participants we can observe their outcomes in the no-program state, but not the outcomes in the program state.

Although it is virtually impossible to measure the program impact for each subject (Heckman, et al., 1999), we can estimate an average impact for a group of subjects. There are two popular parameters in the literature on impact evaluations: the average treatment effect (ATE) and average treatment effect on the treated (ATT). ATE is the expected impact of a program on a person who is randomly assigned to the program. It is equal to the difference in the average outcome of the population between the program state and the no-program state. ATT can be defined as ATE conditional on the program participation. It is equal to the difference in the observed outcome of the participants and their counterfactual outcome if they had not participated in the program. The main difficulty is to estimate the average counterfactual outcomes. If there are concurrent factors that affect outcome and we are unable to net out the impact of these factors from program impact, the counterfactual estimates will be biased. There are a large number of impact evaluation methods, and each method relies on its identification assumption to estimate one or several parameter of the program impacts. Only when the identification assumptions hold, a method can be used to estimate a program impact parameter consistently or unbiasedly.

This paper presents an overview of the most popular impact evaluation methods which are used to measure the ATE and ATT of programs. In addition to a randomization-based method in which participants are selected randomly, these methods are categorized into: (1) methods assuming "selection on observable", and (2) methods assuming "selection on unobservable". If the impact of the program of interest is correlated with other factors affecting the population, we need to isolate the program impact. "Selection on observable" methods are based on an assumption that we can observe all these correlated factors. In contrast, if we are not able to observe all the correlated factors, we need to resort to "selection on unobservable" methods. The paper discusses the identification assumptions and estimation strategy of each method using a unified framework of counterfactuals and a two-equation model.

There are a large number of studies on impact evaluations, both theoretical and practical. Impact evaluation methods are reviewed and discussed in several studies such as Moffitt (1991), Heckman et al. (1999), Blundell & Costa-Dias (2009), Imbens & Wooldridge (2009), Asian Development Bank (2011). Impact evaluation methods are also emphasized in several econometrics book such as Wooldridge (2001) and Angrist & Pischke (2009). Compared with previous studies, this paper is differentiated in two facets. Firstly, we will focus discussion on the identification assumptions and estimation strategy of the most widely-used methods in impact evaluation using a unified framework of counterfactuals and a two-equation model. Methods are compared based on their difference in identification assumptions and their pros and cons in application. For a given program, readers will be able to select relevant impact evaluation methods if there is information on the selection process of participants and data availability for impact evaluation. Secondly, we try to discuss impact evaluation methods using simple mathematic notations so that the discussion can be understood with basic knowledge of statistics or econometrics. For simplicity we focus on identification assumptions of impact evaluation methods instead of assumptions required for specific econometrics estimators.

The paper is structured into six sections. Section 2 gives an overview of the problems in program impact. Section 3 illustrates how random selection can solve these problems. Next, sections 4 and 5 introduce methods relying on selection of observables and methods relying on selection of unobservable, respectively. Finally, section 6 concludes.

2 Problems in program impact evaluation

2.1 Framework of program impact evaluation

The main objective of impact evaluation of a program is to assess the extent to which the program has changed outcomes for subjects. In other words, impact of the program on the subjects is measured by the change in welfare outcome that is attributed only to the program. In the literature on impact evaluation, a broader term "treatment" is sometimes used instead of program/project to refer to intervention whose impact is evaluated.

To make the definition of impact evaluation more explicit, suppose that there is a program assigned to some people in a population *P*. For simplicity, let's assume that there is a single program, and denote by *D* the binary variable of participation in the program, i.e. $D=1$ if she/he participates in the program, and $D=0$ otherwise. Further let *Y* denote the observed value of the outcome. This variable can receive two values depending on the participation variable, i.e. $Y=Y_1$ if $D=1$, and $Y=Y_0$ if $D=0$.[1] These outcomes are considered at a point in time or over a period of time after the program is implemented.

The impact of the program on the outcome of person i is measured by:

$$\Delta_i = Y_{i1} - Y_{i0}, \tag{2.1}$$

which is the difference in outcome between the program state and the no-program state. The problem is that we cannot observe both terms in equation (2.1) for the same person. For those who participated in the program, we can observe only Y_1, and for those who did not participate in the program we can observe only Y_0.

It is practically impossible to estimate the program impact for each person (Heckman, et al., 1999), because we cannot know the counterfactual outcome exactly. Program impact can, however, be estimated for a group of people. In the literature on program impact evaluation, two popular parameters are the Average Treatment Effect (ATE), and the Average Treatment Effect on the Treated (ATT).

ATE is the expected impact of the program on a person who is randomly selected and assigned to the program. It is defined as:

$$ATE = E(\Delta) = E(Y_1 - Y_0) = E(Y_1) - E(Y_0). \tag{2.2}$$

Most programs are targeted to certain subjects. The important question is the program impact on those who participated in the program. The expected treatment effect on the participants is equal to:

$$ATT = E(\Delta \mid D=1) = E(Y_1 - Y_0 \mid D=1) = E(Y_1 \mid D=1) - E(Y_0 \mid D=1). \tag{2.3}$$

[1] *Y* can be a vector of outcomes, but for simplicity let's consider a single outcome of interest.

Except for the case of randomized programs that is discussed in section 3, ATE and ATT are, in general, different from each other, since program participation often depends on the potential outcomes, and as a result $E(Y_1) \neq E(Y_1 \mid D=1)$, and $E(Y_0) \neq E(Y_0 \mid D=1)$. To see this, equation (2.2) can be rewritten as:

$$\begin{aligned} ATE = E(Y_1) - E(Y_0) &= \left[E(Y_1 \mid D=1)\Pr(D=1) + E(Y_1 \mid D=0)\Pr(D=0)\right] \\ &\quad - \left[E(Y_0 \mid D=1)\Pr(D=1) + E(Y_0 \mid D=0)\Pr(D=0)\right] \\ &= \left\{\left[E(Y_1 \mid D=1) - E(Y_0 \mid D=1)\right]\Pr(D=1)\right\} \\ &\quad + \left\{\left[E(Y_1 \mid D=0) - E(Y_0 \mid D=0)\right]\Pr(D=0)\right\}, \end{aligned} \tag{2.4}$$

where $Pr(D=1)$ and $Pr(D=0)$ are the proportions of participants and non-participants of the program, respectively.

Define the average treatment effect on the non-treated (ATNT) as:

$$ANTT = E(Y_1 \mid D=0) - E(Y_0 \mid D=0). \tag{2.5}$$

This parameter can be explained as the effect that the non-participants would have gained if they had participated in the program. Then, ATE can be written as follows:

$$ATE = ATT \Pr(D=1) + ATNT \Pr(D=0). \tag{2.6}$$

Estimation of ATE and ATT is not straightforward, since there are some components that cannot be observed directly. The counterfactual terms $E(Y_1 \mid D=0)$ and $E(Y_0 \mid D=1)$ are not observed. $E(Y_1 \mid D=0)$ is the expected outcome of the participants had they not participated in the program, while $E(Y_0 \mid D=1)$ is the expected outcome of non-participants had they participated in the program. Thus the estimation of *ATE* and *ATT* is not straightforward, and the different methods discussed in this study provide estimates under certain assumptions on how the program is assigned to the population and how the outcome is determined.

Note that we can allow program impact to vary across a vector of observed variables, *X*, since we might be interested in the program impact on certain groups that are specified by the characteristics, *X*. The so-called conditional parameters are expressed as follows:

$$ATE_{(X)} = E(\Delta \mid X) = E(Y_1 \mid X) - E(Y_0 \mid X), \tag{2.7}$$

and:

$$ATT_{(X)} = E(\Delta \mid X, D=1) = E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=1). \tag{2.8}$$

If we denote by $ATNT_{(X)}$ the ATNT conditional on *X*:

$$ATNT_{(X)} = E(\Delta \mid X, D=0) = E(Y_1 \mid X, D=0) - E(Y_0 \mid X, D=0), \tag{2.9}$$

then, similar to (2.7):

$$ATE_{(X)} = ATT_{(X)} \Pr(D=1 \mid X) + ATNT_{(X)} \Pr(D=0 \mid X), \quad (2.10)$$

where $\Pr(D=1 \mid X)$ and $\Pr(D=0 \mid X)$ are the proportion of the participants and non-participants given the *X* variables, respectively.

In the following discussion, we will focus on the conditional parameters - $ATE_{(X)}$ and $ATT_{(X)}$ - since if they are identified, the unconditional parameters - ATE and ATT - can also be identified:

$$ATE = \int_X ATE_{(X)} dF(X), \quad (2.11)$$

$$ATT = \int_{X|D=1} ATT_{(X)} dF(X \mid D=1). \quad (2.12)$$

2.2 Econometric framework of program impact evaluation

A popular way to discuss assumptions of impact evaluation methods is to use the model of two outcome equations (Heckman et al., 1999), in which potential outcomes Y_0 and Y_1 are expressed as functions of individual characteristics (conditioning variables), *X*:[2]

$$Y_0 = \alpha_0 + X\beta_0 + \varepsilon_0 \quad (2.13)$$

$$Y_1 = \alpha_1 + X\beta_1 + \varepsilon_1 \quad (2.14)$$

Y_0 and Y_1 can be any functions of *X*, not necessarily linearly or parametrically specified, and all the identification strategies presented in this paper are still valid. However, to illustrate ideas and links with the traditional linear regression framework, we assume linearity.

For simplicity and identification of program impact in some parametric regressions, we require *X* to be exogenous in the potential outcome equations.

Assumption 2.1: $E(\varepsilon_0 \mid X) = E(\varepsilon_1 \mid X) = 0$ (A.2.1)

In addition, two additional assumptions are needed for the validity of the micro-approach of program impact evaluation. The first assumption is common in the partial equilibrium approach, and required in the literature on program impact evaluation. This assumption is called the stable unit treatment assumption.

Assumption 2.2: $Y_i \perp D_j \ \ \forall i, j$, i.e., realized (observed) outcome of individual *i*, Y_i, is independent of the program status of individual

j, D_j. (A.2.2)

This assumption implies that there is no spill-over effect of the program. In other words, an individual's participation in the program does not affect the outcome of other people.[3]

The second assumption is implicit in the two equation model. Writing the same *X* variables in the two equations (2.15) and (2.16) means that for each person the

[2] For simplicity, subscript i is dropped.

[3] For more detailed discussion on general equilibrium approach in impact evaluation, see, e.g., Heckman, et al. (1999), and Heckman, et al. (1998b)

status of program participation (treatment status) does not affect *X*. Formally speaking, once conditional on potential outcomes, *X* are independent of *D*.

Assumption 2.3:[4] $X \perp D \mid Y_0, Y_1$ (A.2.3)

This assumption does not mean that *X* is uncorrelated with *D*, but that *X* is uncorrelated with *D* given the potential outcomes. Under this assumption *D* does not affect *X* once conditioning on the potential outcomes. Although this assumption is not an indispensable condition to identify program impact, it is maintained for simplicity. If *D* affects *X*, it is much more complex to capture the true impact of program. In the following discussions of different methods in impact evaluation, assumptions 2.1, 2.2 and 2.3 are implicitly assumed to hold.

In the two-equation framework, the parameters of interest for impact evaluation are expressed as follows:

$$\begin{aligned} ATE_{(X)} &= E(Y_1 \mid X) - E(Y_0 \mid X) \\ &= E[\alpha_1 + X\beta_1 + \varepsilon_1 \mid X] - E[\alpha_0 + X\beta_0 + \varepsilon_0 \mid X] \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) \end{aligned} \quad (2.15)$$

and,

$$\begin{aligned} ATT_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=1) \\ &= E[\alpha_1 + X\beta_1 + \varepsilon_1 \mid X, D=1] - E[\alpha_0 + X\beta_0 + \varepsilon_0 \mid X, D=1] \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 - \varepsilon_0 \mid X, D=1). \end{aligned} \quad (2.16)$$

It should be noted that even if coefficients $\alpha_0, \alpha_1, \beta_0, \beta_1$ can be estimated, $ATT_{(X)}$ still includes an unobservable term $E(\varepsilon_1 - \varepsilon_0 \mid X, D=1)$, while $ATE_{(X)}$ does not. To identify $ATT_{(X)}$, in some cases, we need the following additional assumption:

Assumption 2.4: $E(\varepsilon_0 \mid X, D=1) = E(\varepsilon_1 \mid X, D=1)$ (A.2.4)

This assumption states that given *X*, the expectation of the unobserved variables for the participants is the same regardless of the program so that the unobserved term in (2.18) vanishes. It is worth noting that assumption (A.2.4) does not mean the expectation of the error terms conditional on all the *X* variables. Instead, this assumption is required for some variables of *X* that we are interested in the conditional parameters. There might be many explanatory variables *X*, but we are often interested in $ATE_{(X)}$ and $ATT_{(X)}$ conditional on a certain number of variables in *X*, not all *X*. For example, suppose if we want to estimate impacts of a program on income for different age groups, we need (A.2.4) for age only, i.e., $E(\varepsilon_0 \mid age, D=1) = E(\varepsilon_1 \mid age, D=1)$.

To link the counterfactual data with the observed data, substitute (2.13) and (2.14) into the switching model in (2.3). This results in:

[4] Another expression for conditional independence $f(X \mid Y_0, Y_1) = f(X \mid D, Y_0, Y_1)$, where f(.) is conditional density of *X*. For discussion on conditional independence, see, e.g., Dawid (1979).

$$Y = D(\alpha_1 + X\beta_1 + \varepsilon_1) + (1-D)(\alpha_0 + X\beta_0 + \varepsilon_0)$$
$$= \alpha_0 + X\beta_0 + D\left[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + (\varepsilon_1 - \varepsilon_0)\right] + \varepsilon_0. \quad (2.17)$$

Equation (2.17) is a rather general model of program impact, in which the program impact is measured by the coefficient of variable *D* varies across subjects. This coefficient depends on both observable and unobservable variables, *X* and ε. It can also be correlated with *D* if *D* is correlated with *X* and ε. This is a random coefficient model in which the coefficient is correlated with observed and unobserved characteristics variables.

3. Method based on randomized design

3.1. Impact measurement of randomized programs

The randomized design has been an emerging method which can provide the ideal estimator of impact evaluation with robust internal validity (Duflo et al., 2008; Abhijit et al., 2008; Banerjee & Duflo, 2011). In a simple experimental design, a program is assigned randomly to subjects, and those who are assigned the program are willing to participate. In this case, program assignment *D* is said to be independent of the potential outcomes $Y_0$and Y_1. We can state this condition as an assumption.

Assumption 3.1: $Y_0, Y_1 \perp D$ (A.3.1)

Under assumption (A.3.1), parameters $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified.[5] The program impact is estimated simply by comparing the mean outcome between the participants and non-participants. When we have post-program data from a representative sample on participants and non-participants in a randomized program, we can use sample mean of outcomes for treatment and control group to estimate ATE, ATT, and their conditional version $ATE_{(X)}$ and $ATT_{(X)}$.

In reality, we are often interested in impact of a program that is targeted at specific subjects. For example, poverty reduction programs aim to provide the poor with support to get rid of poverty. Vocational training programs are targeted at the unemployed. The program is not assigned randomly to people in the population. In this case, experimental designs can be used to evaluate the impact of the targeted program.

A randomization design or experiment is conducted by choosing a group of people who are willing to participate in the experiment. Denote by D^* the variable indicating the experiment participation. $D^* = 1$ for those in the experiment, and $D^* = 0$ otherwise. Among people with $D^* = 1$, we randomly select people for program participation. Denote R as a variable that $R = 1$ for the participants, and $R = 0$ for non-participants in the experiment. The participants are called the treatment group, while the non-participants (among those in the experiment) are called the control group (or comparison group).

The randomization of program among those in the experiment is stated formally as follows:

[5] Assumption (A.2.1) is made for all methods in impact evaluation.

Assumption 3.2:[6] $Y_0, Y_1 \perp R \mid D^* = 1$ (A.3.2)

To estimate both $ATE_{(X)}$ and $ATT_{(X)}$, we need an additional assumption:

Assumption 3.3: $E(Y_1 \mid X, D = 0) = E(Y_1 \mid X, D = 1) = E(Y_1 \mid X, D^* = 1)$
$E(Y_0 \mid X, D = 0) = E(Y_0 \mid X, D = 1) = E(Y_0 \mid X, D^* = 1)$ (A.3.3)

That is, once conditional on *X*, the expected outcome of those in the experiment is the same as the expected outcome of those not participating in the experiment. It is implied that people who participate in the experiment are similar to those in the reality once conditional on *X*.

Proposition 3.1: $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified under assumptions (A.3.2) and (A.3.3).

Proof:

Under (A.3.2) and (A.3.3), ATT is identified:

$$\begin{aligned} ATT_{(X)} &= E(Y_1 \mid X, D = 1) - E(Y_0 \mid X, D = 1) \\ &= E(Y_1 \mid X, D^* = 1) - E(Y_0 \mid X, D^* = 1) \\ &= E(Y_1 \mid X, D^* = 1, R = 1) - E(Y_0 \mid X, D^* = 1, R = 0), \end{aligned} \quad (3.1)$$

and similarly, the average treatment effect on the non-treated (ATNT) is the same:

$$\begin{aligned} ATNT_{(X)} &= E(Y_1 \mid X, D = 0) - E(Y_0 \mid X, D = 0) \\ &= E(Y_1 \mid X, D^* = 1) - E(Y_0 \mid X, D^* = 1) \\ &= E(Y_1 \mid X, D^* = 1, R = 1) - E(Y_0 \mid X, D^* = 1, R = 0). \end{aligned} \quad (3.2)$$

Thus, the $ATE_{(X)}$ is identified and the same as $ATT_{(X)}$ due to (2.10).

As a result, (3.1) is the unbiased estimator of $ATT_{(X)}$ and $ATE_{(X)}$. We simple calculate the difference in the mean outcome between the participants and non-participants of the program among those attending the experiment. Once the conditional parameters are identified, the conditional parameters are also identified because of (2.11) and (2.12).

3.2. Advantages and disadvantages of the method based on randomization

There is no controversy that among methods of program impact evaluation, the method that is based on randomization of the program produces the most reliable results. Another advantage of the method is the ease in explaining its results to program designers and policy makers, who often do not have much knowledge of statistics and econometrics. The randomized–program method, however, suffers from several drawbacks. Firstly, it is hardly to randomize a program which is targeted at a specific group due to issues of ethics and politics. Randomization of a program means exclusion of some eligible people from the program. It is unfair to

[6] Assumption (A.3.2) states that the selection of participants among the experimental people is independent of the potential outcomes. In fact we only need a weaker version to identify ATT:
$E(Y_1 \mid D^* = 1, R = 1) = E(Y_1 \mid D^* = 1, R = 1)$ and $E(Y_0 \mid D^* = 1) = E(Y_0 \mid D^* = 1, R = 0)$.
However this assumption is difficult to interpreter. Thus we mention the assumption (A.3.2) in discussing the identification of the program impact.

deny (or delay) a program that provides supports such as health care or education for some eligible people.

Secondly, the implementation and evaluation of a socioeconomic program that is based on randomization is often expensive. Subjects are scattered in the population, which increases the cost of program administration and data collection for impact evaluation.

Thirdly, there can be some factors that bias the estimates from randomization-based evaluation. These factors invalidate the key identification assumption (A.3.1), $D \perp Y_0, Y_1$. Two problems that are widely mentioned are attrition and substitution effects.

Attrition means that some people in the treatment group quit the program during implementation. As a result, their observed outcome is not the potential outcome in the presence of the program, Y_1. If this drop-out is random, there is no concern about this problem since the randomization feature remains preserved. If the attrition is not random but correlated with some characteristics of the drop-outs, the remaining subjects in the treatment group who actually take the program will be systematically different from the subjects in the control group. In other words, there is self-selection into the program of the participants, which is dealt with by the alternative methods discussed in the following sections. The mean difference in outcome between the treatment and control group is not an estimator of the program impact, but an estimator of "the mean effect of the offer of treatment" (Heckman, et al., 1999).

The substitution effect means that some people in the control group might try to get access to programs that are similar to the program to be evaluated. The substitution programs can contaminate the outcome of the control group. It is implied that if the program had not been implemented, the participants would have taken other similar programs. The mean difference in outcome between the control and treatment groups reflects "the mean incremental effect of the program relative to the world in which it does not exist" (Heckman, et al., 1999). To truly capture the program impact, we need to have information on impacts of the substituted programs, and subtract them from the outcome of the control group to estimate the potential outcome of the treatment group in the absence of the program.

Finally, a randomized program that is used for impact-evaluation purposes is often a pilot program, and the impact of the pilot program can be far from the impact of the program when it is implemented in reality. A pilot program is often smaller and more easily administered.

4 Methods assuming selection on observables

4.1 Selection bias and conditional independence assumption

When a program is not assigned randomly, the potential outcomes of the participants will be different from those of non-participants. Assumption (A.3.1) no longer holds, and simple comparison of mean outcomes between participants and non-participants contain the selection bias. To see the selection bias in estimating the average treatment effect $ATE_{(X)}$ conditioning on X, rewrite the formula of $ATE_{(X)}$:

$$\begin{aligned} ATE_{(X)} &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 - \varepsilon_0 \mid X) \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) \\ &+ \{[\Pr(D=1 \mid X)E(\varepsilon_1 \mid X, D=1) + \Pr(D=0 \mid X)E(\varepsilon_1 \mid X, D=0)] \\ &- [\Pr(D=1 \mid X)E(\varepsilon_0 \mid X, D=1) + \Pr(D=0 \mid X)E(\varepsilon_0 \mid X, D=0)]\}. \end{aligned} \tag{4.1}$$

When we use the following estimator:

$$\begin{aligned} \hat{ATE}_{(X)} &= E(Y_1 \mid X, D=1) - E(Y_0 \mid X, D=0) \\ &= E(\alpha_1 + X\beta_1 + \varepsilon_1 \mid X, D=1) - E(\alpha_0 + X\beta_0 + \varepsilon_0 \mid X, D=0) \\ &= (\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0) + E(\varepsilon_1 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0), \end{aligned} \tag{4.2}$$

the bias is equal to:

$$\begin{aligned} \hat{ATE}_{(X)} - ATE_{(X)} &= \left[E(\varepsilon_1 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0)\right] \\ &\quad - \{\left[\Pr(D=1 \mid X)E(\varepsilon_1 \mid X, D=1) + \Pr(D=0 \mid X)E(\varepsilon_1 \mid X, D=0)\right] \\ &\quad - \left[\Pr(D=1 \mid X)E(\varepsilon_0 \mid X, D=1) + \Pr(D=0 \mid X)E(\varepsilon_0 \mid X, D=0)\right]\} \\ &= \Pr(D=0 \mid X)\left[E(\varepsilon_1 \mid X, D=1) - E(\varepsilon_1 \mid X, D=0)\right] \\ &\quad + \Pr(D=1 \mid X)\left[E(\varepsilon_0 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0)\right]. \end{aligned} \tag{4.3}$$

Even though X are controlled for, selection bias in estimating $ATE_{(X)}$ can arise if the conditional expectation of unobserved variables in potential outcomes, ε_0 and ε_1, is different for the participants and non-participants.

Similarly, if we use the same estimator in (4.2) for $ATT_{(X)}$, the selection bias will be:

$$\hat{ATT}_{(X)} - ATT_{(X)} = E(\varepsilon_0 \mid X, D=1) - E(\varepsilon_0 \mid X, D=0). \tag{4.4}$$

The selection bias stems from the difference in the conditional expectation of unobserved variables, ε_0, between the participants and non-participants.[7]

One intuitive way to avoid the selection biases, (4.3) and (4.4), in estimating $ATE_{(X)}$ and $ATE_{(X)}$ is to invoke assumptions so that the selection biases are equal to zero. The assumption on "selection on observable" assumes that one is able to observe all variables that affect both the program selection and potential outcomes so that once conditioned on these variables, the potential outcomes Y_0 and Y_1 are independent of the program assignment. In Rosenbaum & Rubin (1983), this assumption is called ignorability of treatment or conditional independence. Formally, it is written as:

Assumption 4.1: $Y_0, Y_1 \perp D \mid X$ (A.4.1)

Assumption (A.4.1) can be considered as a conditional version of assumption (A.3.1). Once we have control for X, the assignment of the program becomes randomized. A corollary of assumption (A.4.1) is that the error terms in the potential outcomes is also independent of D given X, i.e.:

$$\varepsilon_0, \varepsilon_1 \perp D \mid X. \tag{4.5}$$

[7] If one has data before and after a program, they sometimes use the before and after estimator to estimate the program impact. The bias is equal to $E(Y_{0B} \mid D=1) - E(Y_{0A} \mid D=1)$, where $E(Y_{0B} \mid D=1)$ and $E(Y_{0A} \mid D=1)$ are the expectation of participants' outcome in the state of no program before and after the program, respectively. The assumption is valid if there is no change in the participants' outcome during the program implementation if they had not participated. Intuitively, this assumption seems plausible in short time, but might be unreasonable in long time.

Under condition (4.5), we have (Dawid, 1979):

$$E(\varepsilon_0 \mid X, D=0) = E(\varepsilon_0 \mid X, D=1), \quad (4.6)$$

$$E(\varepsilon_1 \mid X, D=0) = E(\varepsilon_1 \mid X, D=1). \quad (4.7)$$

As a result of equation (4.6) and (4.7), the selection biases given in (4.3) and (4.4) are equal to zero. $ATE_{(X)}$ and $ATT_{(X)}$ are identified, and so are ATE and ATT.

In addition, assumption (4.5) results in:

$$E(\varepsilon_1 - \varepsilon_0 \mid X, D=1) = E(\varepsilon_1 - \varepsilon_0 \mid X) = 0. \quad (4.8)$$

Hence, $ATE_{(X)}$ is equal to $ATT_{(X)}$. Assumption (A.4.1) is the key assumption for identifying program impacts that "selection on observables" methods rely on. This does not mean that we have to observe all information on the program selection, i.e. *D* is deterministic, but it implies that all the *X* variables that make *D* correlated with Y_0 and Y_1 are observed. Three widely-used sets of methods that use this assumption are presented in this paper, namely regression methods, matching methods, regression discontinuity. All these methods can be conducted using single cross section data.

4.2. Regression methods assuming selection on observables

For simplicity we maintain the assumption of linearity in outcome equations for this section. Next we will discuss the case of nonlinear functions of potential outcomes.

Proposition 4.1: Given assumptions (A.4.1), OLS regression produces unbiased estimators of $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT.

Proof: The observed outcome is as follows:

$$Y = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + [D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0] \quad (4.9)$$

The proof is now similar to the proof of Proposition 3.2. The error term has the following property:

$$E[D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0 \mid X, D] = DE(\varepsilon_1 - \varepsilon_0 \mid X, D) + E(\varepsilon_0 \mid X, D) = E(\varepsilon_0 \mid X) = 0 \quad (4.10)$$

Under assumption (A.4.1), $ATE_{(X)}$ and $ATT_{(X)}$ are the same, and the estimators of these conditional parameters are:

$$\hat{ATE}_{(X)} = \hat{ATT}_{(X)} = (\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\beta}_1 - \hat{\beta}_0)X. \quad (4.11)$$

ATE and ATT are identified simply by taking the expectation of $ATE_{(X)}$ and $ATT_{(X)}$ over the distribution of *X* for the whole population, and the distribution of *X* for the participant population, respectively.

The regression methods have the advantage of simple implementation, but also have three main drawbacks. Firstly, they impose a specific functional form on the relation between outcome and conditioning variables and the program participation variable. Secondly, because of the functional form, the OLS regression can have, making the estimator of the program impact inefficient will be inefficient if the parametric regressions are plagued by problems of multicollinearity and

heteroscedasticity. Finally, the method relies on the assumption of program selection based on the observable variables. This assumption is strong.

4.3 Matching methods

Identification assumptions

There is a large amount of literature on matching methods of impact evaluation. Important contributions in this area can be found in studies such as Rubin (1977; 1979; 1980), Rosenbaum & Rubin (1983; 1985a), and Heckman, et al. (1997b). The matching method can be used to estimate the two program impact parameters, ATE and ATT under the conditional independence assumption (A.4.1). The basic idea of the matching method is to find a control group (also called comparison group) that has the same (or at least similar) distribution of X as the treatment group. By doing so, we have controlled for the difference in X between the participants and non-participants. The potential outcomes of the control and treatment group are now independent of the program selection. The difference in outcome of the control group and the treatment group then can be attributed to the program impact.

However for the matching method to be implemented, we must find a control group that is similar to the treatment group but does not participate in the program. This similarity assumption is called common support. If we denote p(X) as the probability of participating in the program for each subject, i.e. $p(X) = P(D=1|X)$, the assumption can be stated formally as follows:

Assumption 4.2: $0 < p(X) < 1$ (A.4.2)

Proposition 4.2: Under assumptions (A.4.1) and (A.4.2), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified by the matching method.

Proof: the proof is straightforward using the conditional independence assumption.

$$ATE_{(X)} = ATT_{(X)} = E(Y_1|X) - E(Y_0|X) = E(Y_1|X, D=1) - E(Y_0|X, D=0) \quad (4.12)$$

Both terms in (4.20) can be observed. In addition, assumption (A.4.2) ensures that there are some participants and non-participants whose values of X are the similar so that we are able to use sample information to estimate (4.24).

ATE and ATT are identified as in (2.13) and (2.14).

Construction of a comparison group

To implement the matching method, we need to find a comparison group for which the conditioning variables are comparable to those of the treatment group. The comparison group is constructed by matching each participant i in the treatment group with one or more non-participants j whose variables X_j are closest to X_i of the participant i. The weighted average outcome of non-participants who are matched with an individual participant i will form the counterfactual outcome for the participant i.

For a participant i, denote n_{ic} as the number of non-participants j who are matched with this participant, and $w(i,j)$ the weight attached to the outcome of each non-participant. These weights are defined non-negative and sum up to 1, i.e. $\sum_{j=1}^{n_{ic}} w(i,j) = 1$.

The estimator of the conditional program parameters is then equal to:

$$\hat{ATE}_{(X)} = \hat{ATT}_{(X)} = \frac{1}{\sum_{i \text{ with } X_i = X} D_i} \left\{ \sum_{i \text{ with } X_i = X} \left[Y_{1i} - \sum_{j=1}^{n_{ic}} w(i,j) Y_{0j} \right] \right\} \quad (4.13)$$

where Y_{1i} and Y_{0j} are the observed outcomes of participant i and non-participant j. ATT is simply the average of differences in outcome between the treatment and comparison group:

$$\hat{ATT} = \frac{1}{n_1} \sum_{i=1}^{n_1} \left[Y_{1i} - \sum_{j=1}^{n_{ic}} w(i,j) Y_{0j} \right] \quad (4.14)$$

where n_1 is the number of the participants in the data sample.

To estimate the ATE, we also need to estimate the effect of non-treatment on the non-treated using an estimator as follows:

$$\hat{ANTT} = \frac{1}{n_2} \sum_{j=1}^{n_2} \left[Y_{0j} - \sum_{i=1}^{n_{jt}} w(j,i) Y_{1i} \right] \quad (4.15)$$

where n_2 is the number of the non-participants in the sample. n_{jt} is the number of participants is matched with a non-participant j, and w(j,i) are weights attached to each participant i in this matching.

Thus using (2.6) the estimator of ATE is expressed as follows:

$$\hat{ATE} = \frac{1}{n_1 + n_2} \left\{ \sum_{i=1}^{n_1} \left[Y_{1i} - \sum_{j=1}^{n_{ic}} w(i,j) Y_{0j} \right] + \sum_{i=1}^{n_2} \left[Y_{0j} - \sum_{i=1}^{n_{jt}} w(j,i) Y_{1i} \right] \right\} \quad (4.16)$$

To this end, there are still two essential issues that have not been discussed. The first is how to select non-participants and participants for matching. The second is how to determine weights w(i,j) among these matched people.

Methods to find a matched sample

Clearly, matched non-participants should have X closest to X of participants. There will be no problem if there is a single conditioning variable X. However X is often a vector of variables, and finding "close" non-participants to match with a participant is not straightforward. In the literature on impact evaluation, there are three widely-used methods to find matched non-participants for a participant (and vice versa matched participants for a non-participant).

The first method is called subclassification of the treatment and control group based on X (see, e.g., Cochran & Chambers, 1965; Cochran, 1968). All participants and non-participants are classified into blocks according to the value of X. This means that subjects in a block have the same value of X. Then non-participants will be matched with participants in each block. However the subclassification becomes difficult when there are many variables X or when some variables of X are continuous or discrete with many values.

The second method is called covariate matching and matches participants with non-participants based on their distance of variables defined on some metric (Rubin, 1979; 1980). Since X can be considered as a vector in a space, the closeness between two sets of X can be defined by a distance metric. A non-participant j will be matched with a participant i if the distance from X_j to X_i is smallest as compared with other non-participants using traditional Euclidean metric

such as the Mahalanobis metric (Rubin, 1979; 1980) or the inversed variance matrix of X (Abadie & Imbens, 2002).[8]

The third way to find the matched sample is the propensity score matching. Since a paper by Rosenbaum & Rubin (1983), matching is often conducted based on the probability of being assigned to the program, which is called the propensity score. Rosenbaum & Rubin (1983) show that if the potential outcomes are independent of the program assignment given X, then they are also independent of the program assignment given the balance score. The balance score is any function of X but finer than $p(X)$, which is the probability of participating in the program (the so-called propensity score). In fact, the propensity score is often selected as the balance score in estimating the program impacts. The propensity score can estimated parametrically or non-parametrically by running a regression of the treatment variable D on the conditioning variables X. Since D is a binary variable, a logit or probit model is often used. Once the propensity score is obtained for all subjects in the sample, non-participants can be matched with participants based on the closeness of the propensity scores[9].

Weighting methods of matched comparisons

Once a metric distance, $d(i,j)$, between a participant i and a non-participant j is defined, one can select methods to weight their outcomes. If each participant is matched with the one non-participant with the minimum value of $d(i,j)$, the weight $w(i,j)$ equals 1 for all pairs of matches. This is called one nearest neighbor matching. When more than one non-participants are matched with each participant (or vice versa), we need some ways to define the weights attached to each non-participant.

A number of methods use equal weights for all matches. N-nearest neighbor matching involves matching each participant with n non-participants whose have the closest distances $d(i,j)$. Each matched non-participant will receive weight $w(i,j) = 1/n$. Caliper matching (see, e.g., Dehejia & Wahba, 1998; Smith & Todd, 2005) uses equal weights for matched subjects whose distance $d(i,j)$ is smaller than a specific value, say 0.05 or 0.1. This criterion aims to ensure the quality of matching. Stratification (interval) matching divides the range of estimated distances into several strata (blocks) of equal ranges. Within each stratum, a participant is matched with all non-participants with equal weights (see, e.g., Dehejia & Wahba, 1998); Smith & Todd, 2005).

However, it could be reasonable to assign different weights to different non-participants depending on metric distances between their covariates and the covariates of the matched participant. This argument motivates some others matching schemes such as kernel, local linear matching (see, e.g., Heckman, et al., 1997b; Smith & Todd, 2005), and matching using weights of inversed propensity score (see, e.g., Hahn, 1998; Hirano, et al., 2002).

The main advantage of the matching method is that it does not rely on a specific functional form of the outcome, thereby avoiding assumptions on functional form. In addition, the matching method emphasizes the problem of common support, thereby avoiding the bias due to extrapolation to non-data region. However, the main limitation of the matching method is that it relies on the strong assumption of conditional mean independence.

4.4. Regression discontinuity design

For the matching method, the assumption on the common support is required to identify the program impacts. When the conditioning variables X are different for

[8] The Mahalanobis metric is presented in Mahalanobis (1936).

[9] The propensity score can also be used instead of X in regressions to estimate program impact (see, e.g., Wooldridge, 2001; Rosenbaum & Rubin, 1985a).

participants and non-participants, we cannot implement matching methods. In other words, if there are some variables X that predict the treatment variable D perfectly, the assumption of common support no longer holds. In Van der Klaauw (2002), it means that there is a conditioning variable S belonging to X such that D equals 1 if and only if S is larger than a specific value $\overline{S}$.[10] For example, social pension is provided for all the elderly above at a given threshold, say 65 years old. People older than 65 receive pensions, while others from 65 and below do not receive pension.

In this case, the assignment of the program is called deterministic. To make this assumption consistent with notation in this paper, we assume that $D=1$ if and only if $X \geq \widetilde{X}$. Then we have:

$$P(D=1 \mid X \geq \widetilde{X}) = 1, \quad (4.17)$$

$$P(D=1 \mid X < \widetilde{X}) = 0. \quad (4.18)$$

Which means that the common support assumption $0 < P(D=1 \mid X) < 1$ is not valid.

We know that the regression method does not require a common support. As a result it can be applied in this context taking into account some important notes. Under the assumption on conditional mean independence, the conditional and unconditional program impact parameters are the same because of:

$$E(Y_0 \mid X, D=1) = E(Y_0 \mid X, D=0), \quad (4.19)$$

$$E(Y_1 \mid X, D=1) = E(Y_1 \mid X, D=0), \quad (4.20)$$

which can be expressed as follows due to (4.17) and (4.18):

$$E(Y_0 \mid X, X \geq \widetilde{X}) = E(Y_0 \mid X, X < \widetilde{X}), \quad (4.21)$$

$$E(Y_1 \mid X, X \geq \widetilde{X}) = E(Y_1 \mid X, X < \widetilde{X}). \quad (4.22)$$

If the potential outcomes are monotonous (as in case of linear function with first-order variables X), (4.21) and (4.22) are obtained only at the point $X = \widetilde{X}$ under a condition that the potential outcome are continuous at this point. Since the potential outcomes are functions of the error terms, we can state this assumption with respect to the error terms.

Assumption 4.3: The conditional means of the error terms $E(\varepsilon_0 \mid X)$, and $E(\varepsilon_1 \mid X)$ are continuous at $\widetilde{X}$. (A.4.3)

Under assumption (A.4.3) the matching method and other non-parametric estimation methods can be used to estimate the program impacts at the mass of $\widetilde{X}$. This is called local treatment effect at $\widetilde{X}$ (see, e.g.,Van der Klaauw, 2002; Hahn, et al., 2001). Linear regression can also be used to estimate the program impact parameters.

[10] Heckman, et al. (1999) presents the case in which $D=1$ only if $S < \overline{S}$. These two cases are similar.

When the program participation is not absolutely deterministic, i.e. there are some subjects who have $X \geq \tilde{X}$ but do not participate in the program, or some other subjects who have $X < \tilde{X}$ but do participate in the program, one can apply fuzzy regression results in which the X variables can be used as an instrument for the program variable at the threshold $\tilde{X}$ (e.g., Imbens & Lemieux, 2008; and Lee & Lemieux, 2010).

5. Methods assuming selection on unobservables

As discussed, the main assumption that the methods of selection on observable rely on is the conditional independence between the potential outcomes and program assignment (or a weaker version of conditional mean independence). This assumption does not hold if there is an unobserved variable affecting both the potential outcome and program participation. For many programs, people decide to participate in a program based on their complex criteria, which are not observed or measured by impact evaluation practitioners. For example, the poor are eligible for micro-credit, but not all of them are willing to take micro-credit. If people who have better business capacity and motivation for high income are more likely to borrow, it's almost impossible to observe and measure these variables. In this case, the 'selection on observable' methods produce biased estimates of the program impact. This section presents three methods that are widely-used in dealing with the problem of "selection on unobservables". The methods include instrumental variable regression, sample selection models, and panel data models.

5.1. Instrumental variables

Program impact identification

If there are unobserved variables affecting both potential outcomes and program participation, the program variable is endogenous in the outcome equation and OLS gives biased estimates. A standard solution to this endogeneity problem is to use one or more instrumental variables for the program assignment variable *D*. An instrumental variable has two properties: (i) it is correlated with program assignment; and (ii) it is uncorrelated with the error term in the potential outcomes.[11]

To illustrate how the instrumental variables method identifies program impact, recall equation (2.17):

$$Y = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + [D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0]. \qquad (5.1)$$

Assumption 5.1: There is at least an instrumental variable *Z* such that:

$Cov(D, Z) \neq 0$,

$$E(\varepsilon_0 \mid Z) = E(\varepsilon_0), \qquad (A.5.1)$$
$$E(\varepsilon_1 \mid Z) = E(\varepsilon_1).$$

Proposition 5.1: Under assumptions (A.2.4) and (A.5.1), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified and estimated by the instrumental variables method.

Proof:

Firstly we show that:

[11] Examples of instrumental variables can be seen in econometrics textbooks such as Wooldridge (2001), Greene (2003) or papers on review of impact evaluation such as Moffitt (1991).

$$Cov\left(\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],Z\right)=0. \tag{5.2}$$

Note that $E(\varepsilon_1-\varepsilon_0\mid D,Z)=E(\varepsilon_1-\varepsilon_0\mid D)=0$ because of (A.2.4) and (A.5.1), hence:

$$\begin{aligned} Cov\left(\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],Z\right) &= Cov\left(\left[D(\varepsilon_1-\varepsilon_0)\right],Z\right)+Cov(\varepsilon_0,Z)\\ &= E\left\{D(\varepsilon_1-\varepsilon_0)-E\left[D(\varepsilon_1-\varepsilon_0)\right]\right\}\left\{Z-E(Z)\right\}\\ &= E\left[DZ(\varepsilon_1-\varepsilon_0)\right]\\ &= 0. \end{aligned} \tag{5.3}$$

Similar, we have:

$$Cov\left(\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],X\right)=0, \tag{5.3}$$

$$Cov\left(\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],XZ\right)=0. \tag{5.4}$$

Then we have the following covariance equations due to (5.2), (5.3) and (5.4):

$$\begin{aligned} Cov(Y,Z) &= Cov\left\{\alpha_0+X\beta_0+D\left[(\alpha_1-\alpha_0)+X(\beta_1-\beta_0)\right]+\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],Z\right\}\\ &= Cov(X,Z)\beta_0+Cov(D,Z)(\alpha_1-\alpha_0)+Cov(XD,Z)(\beta_1-\beta_0), \end{aligned} \tag{5.5}$$

$$\begin{aligned} Cov(Y,X) &= Cov\left\{\alpha_0+X\beta_0+D\left[(\alpha_1-\alpha_0)+X(\beta_1-\beta_0)\right]+\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],X\right\}\\ &= Var(X)\beta_0+Cov(D,X)(\alpha_1-\alpha_0)+Cov(XD,X)(\beta_1-\beta_0), \end{aligned} \tag{5.6}$$

$$\begin{aligned} Cov(Y,XZ) &= Cov\left\{\alpha_0+X\beta_0+D\left[(\alpha_1-\alpha_0)+X(\beta_1-\beta_0)\right]+\left[D(\varepsilon_1-\varepsilon_0)+\varepsilon_0\right],XZ\right\}\\ &= Cov(X,XZ)\beta_0+Cov(D,XZ)(\alpha_1-\alpha_0)+Cov(XD,XZ)(\beta_1-\beta_0) \end{aligned} \tag{5.7}$$

Since the number of unknown parameters is equal to the number of equations, we can identify the parameters in regression model and indentify the conditional and unconditional ATE and ATT.

It should be noted that equation (2.19) includes the interaction between *X* and *D*. Thus it is considered to include endogenous variables *D* and *XD*, and we use instrumental variables *Z* and *XZ* to solve the endogeneity problem. The instrumental variable method is presented above for just-identification, i.e., only one instrumental variable. The case of over-identification in which there are more than one instrumental variable for the treatment variable *D* can be solved easily by applying two-stage least square regression (see, e.g., Wooldridge, 2001).[12]

Local average treatment effect

The instrumental variable method presented in the above section is standard. It requires assumption (A.2.4) to identify program impact. Imbens & Angrist (1994) proposes an another method of instrumental variables that does not rely on assumption (A.2.4) in identifying a so-called local average treatment effect (LATE). The LATE parameter measures the effect of the program on those who change program status due to a change in an instrumental variable *Z*. As *Z* is defined as a policy or a set of policies, one would be interested in impact of a program on those who are included in the program as a result of policy changes.

To formalize the definition, suppose there is an instrumental variable *Z*, whose value changed from $Z=z_0$ to $Z=z_1$. As a result, there are a number of subjects

[12] For example, in the first stage the propensity score is estimated using instrumental variables. Then in the second stage, the predicted propensity score is used as an instrumental variable in the outcome equation.

who changed their status from non-participation to participation in the program. Further denote $D(z,X)$ is the treatment variable D but conditional on $Z = z$ for subjects with X. Then LATE is defined:

$$LATE_{(X,z_0,z_1)} = E\left[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1\right] \tag{5.8}$$

In addition to the condition of instrumental variables (A.5.1), Imbens and Angrist (1994) impose an additional assumption to identify LATE.

Assumption 5.2: For all z and z' of Z, either $D(z,X) \geq D(z',X)$ or $D(z,X) \leq D(z',X)$ for all subjects. (A.5.2)

In other words, if D can be expressed in a latent variable context, in which D = 1 if D^* is greater than zero, and otherwise, then D^* is required to be monotonous in Z. Once conditional on X, any subject should prefer to participate (or quit) the program as the instrument Z changes its value from z to z'.

Proposition 5.2(Imbens and Angrist, 1994)**:** Under assumption (A.5.1) and (A.5.2), LATE is identified as follows:

$$\begin{aligned} LATE_{(X,z_0,z_1)} &= E\left[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1\right] \\ &= \frac{E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0)}{P(D = 1 \mid X, Z = z_1) - P(D = 1 \mid X, Z = z_0)}, \end{aligned} \tag{5.9}$$

where Y is the observed outcome, and the denominator is different from zero.

Proof: We have:

$$\begin{aligned} E(Y \mid X, Z = z_0) &= E\left\{(Y_1 D(z_0, X) + \left[(1 - D(z_0, X)Y_0\right] \mid X, Z = z_0\right\} \\ &= D(z_0, X)E(Y_1 \mid X) + \left[1 - D(z_0, X)\right]E(Y_0 \mid X), \end{aligned} \tag{5.10}$$

$$\begin{aligned} E(Y \mid X, Z = z_1) &= E\left\{(Y_1 D(z_1, X) + \left[(1 - D(z_1, X)Y_0\right] \mid X, Z = z_1\right\} \\ &= D(z_1, X)E(Y_1 \mid X) + \left[1 - D(z_1, X)\right]E(Y_0 \mid X). \end{aligned} \tag{5.11}$$

Subtract (5.10) from (5.11), we get:

$$\begin{aligned} &E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0) \\ &= \left[D(z_1, X) - D(z_0, X)\right]E(Y_1 - Y_0 \mid X) \\ &= E\left[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1\right]P\left[D(z_1, X) - D(z_0, X) = 1\right] \\ &\quad + E\left[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = -1\right]P\left[D(z_1, X) - D(z_0, X) = -1\right] \\ &= E\left[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1\right]P\left[D(z_1, X) - D(z_0, X) = 1\right]. \end{aligned} \tag{5.12}$$

The last line results from assumption (A.5.2) that there is no person who quits the program due to the change in Z from z_0 to z_1.

Hence:

$$\begin{aligned} E\left[Y_1 - Y_0 \mid X, D(z_1, X) - D(z_0, X) = 1\right] &= \frac{E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0)}{P\left[D(z_1, X) - D(z_0, X) = 1\right]} \\ &= \frac{E(Y \mid X, Z = z_1) - E(Y \mid X, Z = z_0)}{P(D = 1 \mid X, Z = z_1) - P(D = 1 \mid X, Z = z_0)}. \end{aligned} \tag{5.13}$$

The unconditional LATE is identified by taking the expectation of (5.9) over *X*.

Finally, it should be noted that *Z* can be a vector of instrumental variables, and LATE is defined as the program impact on those whose participate in the program due to a change in a set of program policies.

The main advantage of the instrumental variable method is that it allows for the program selection based on unobservable. However, the main problem in this method is to find good instrumental variables. A variable that is correlated with the program selection is often correlated with outcomes and error terms in the potential outcome equations. Using an invalid instrumental variable that does not satisfy the instrument conditions will lead to biased and inconsistent estimates of the program impacts. In contrast, a variable that is uncorrelated with the error terms can be very weakly correlated with the program selection. Weak instruments can result in problems of large standard errors and biased estimates (Staiger & Jame, 1997).

5.2. Sample selection models

Impacts of a program can be identified using a sample selection model (Heckman, 1978). Recall that we cannot run regression of the potential outcomes using sample data in the presence of the selection bias because of the non-random missing data. For example, in the equation of Y_0 there is no data on the dependent variable for those who participated in the program. This is similar to the case of the censored dependent variable model, in which the dependent variables is censored according a selection mechanism. Under assumptions on distribution between the error term in the program selection and the error terms in the potential outcome equations, we can estimate coefficients in the potential outcomes consistently. Let's write the impact evaluation model again:

The potential outcomes:

$$Y_0 = \alpha_0 + X\beta_0 + \varepsilon_0,$$
$$Y_1 = \alpha_1 + X\beta_1 + \varepsilon_1,$$

and the outcome that we observe is:

$$Y = DY_1 + (1-D)Y_0 = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + [D(\varepsilon_1 - \varepsilon_0) + \varepsilon_0],$$

where *D* is determined by the following framework:

$$D^* = \theta W + v,$$
$$D = 1 \text{ if } D^* > 0,$$
$$D = 0 \text{ otherwise.}$$

$ATE_{(X)}$ and $ATT_{(X)}$ can be estimated if we are able to get unbiased estimators of $(\alpha_1 - \alpha_0)$, and $(\beta_1 - \beta_0)$, and the term, $E(\varepsilon_1 - \varepsilon_0 \mid X, D = 1)$.

Assumption 5.3: The error term v in the program participation equation and each of the error terms ε_0, ε_1 in the potential outcome equations follows the following bivariate normal distributions:

$$(v, \varepsilon_0) \sim N_2(0,0,1,\sigma_{\varepsilon_0}, \rho_0)$$
$$(v, \varepsilon_1) \sim N_2(0,0,1,\sigma_{\varepsilon_1}, \rho_1) \quad \text{(A.5.3)}$$

Proposition 5.3: Under assumptions (A.5.3), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified.

Proof:

We have the conditional expectation of the observed outcome in equation (2.17):

$$E(Y \mid X, D) = \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] + E\{D[(\varepsilon_1 - \varepsilon_0)] + \varepsilon_0 | X, D\} \tag{5.14}$$

in which:

$$\begin{aligned} E\{D[(\varepsilon_1 - \varepsilon_0)] + \varepsilon_0 | X, D\} &= DE(\varepsilon_1 - \varepsilon_0 \mid X, D) + E(\varepsilon_0 \mid X, D) \\ &= E(\varepsilon_1 \mid X, D=1)P(D=1 \mid X) + E(\varepsilon_0 \mid X, D=0)P(D=0 \mid X) \\ &= E(\varepsilon_1 \mid X, v > -\theta W)P(D=1 \mid X) + E(\varepsilon_0 \mid X, v \le -\theta W)P(D=0 \mid X) \\ &= \left\{E(\varepsilon_1 \mid X) + \rho_1\sigma_{\varepsilon_1}\frac{\phi(\theta W)}{\Phi(\theta W)}\right\}P(D=1 \mid X) + \left\{E(\varepsilon_0 \mid X) + \rho_0\sigma_{\varepsilon_0}\frac{-\phi(\theta W)}{1-\Phi(\theta W)}\right\}P(D=0 \mid X) \\ &= \left[\rho_1\sigma_{\varepsilon_1}\frac{\phi(\theta W)}{\Phi(\theta W)}\right]P(D=1 \mid X) - \left[\rho_0\sigma_{\varepsilon_0}\frac{\varphi(\theta W)}{1-\Phi(\theta W)}\right]P(D=0 \mid X), \end{aligned} \tag{5.15}$$

where the fourth lines results from the definition of the truncated distribution (see, e.g., Greene, 2003). $\phi(.)$ and $\Phi(.)$ are the probability density function and the cumulative probability function of the standard normal distribution, respectively.

Hence (5.14) has the form:

$$\begin{aligned} Y = {} & \alpha_0 + X\beta_0 + D[(\alpha_1 - \alpha_0) + X(\beta_1 - \beta_0)] \\ & + \left[\rho_1\sigma_{\varepsilon_1}\frac{\phi(\theta W)}{\Phi(\theta W)}\right]P(D=1 \mid X) - \left[\rho_0\sigma_{\varepsilon_0}\frac{\varphi(\theta W)}{1-\Phi(\theta W)}\right][1 - P(D=1 \mid X)] + u, \end{aligned} \tag{5.16}$$

where u is an error term. (5.16) can be estimated by OLS or maximum likelihood methods. Estimates of θ are obtained from estimation of the program selection equation, while $P(D=1 \mid X)$ is the propensity score that can be estimated parametrically or non-parametrically.

To identify $ATT_{(X)}$, we need the estimation the term $E(\varepsilon_1 - \varepsilon_0 \mid X, D=1)$, which is equal to:

$$\begin{aligned} E(\varepsilon_1 - \varepsilon_0 \mid X, D=1) &= E(\varepsilon_1 \mid X, v > -\theta W) - E(\varepsilon_0 \mid X, v > -\theta W) \\ &= \left(\rho_1\sigma_{\varepsilon_1} - \rho_0\sigma_{\varepsilon_0}\right)\frac{\phi(\theta W)}{\Phi(\theta W)}, \end{aligned} \tag{5.17}$$

in which $\rho_1\sigma_{\varepsilon_1}$ and $\rho_0\sigma_{\varepsilon_0}$ are estimated from (5.16).

Although there is no strict requirement of exclusion restriction, i.e. at least an instrumental variable included in W, such an instrumental variable should be included in W to avoid high multicollearity in (5.16). In addition, if we are able to find instrumental variables in W, the expectation of the error terms conditional on X and D can be estimated semi-parametrically or non-parametrically without assumption on the bivariate normal distribution of the error terms (see, e.g., Heckman, 1990; Powell, 1994).

Similar to the method of instrumental variables, the main advantage of the sample selection method is that it allows for selection of a program based on unobservable. In addition, it is robust to heterogeneous impacts of the program. However, the main problem in this method is that it requires the assumption on the functional form of the join distribution of the error terms in the selection equation and the potential outcome equations. In addition, a good instrumental variable is often needed to get efficient estimators of the program impact.

5.3. Panel data methods

In impact evaluation of many programs, baseline and endline surveys are conducted. When longitudinal data or panel data on the participants and non-participants in a program before and after the program implementation are available, we can get unbiased estimators of program impacts which allow for "selection on time-invariant unobservable". Methods discussed in this section are based on the panel data at two points of time, since this type of data are the most popular. For the two-period panel data, the first-different regression is also the same as the fixed-effects regression. This method is easily applied to the case of panel data with more than two periods.

First-difference method

To illustrate how the method identifies the program impact, let's write the model of the outcome before the program implementation as follows:

$$Y_{0B} = \alpha_{0B} + X_B\beta_{0B} + \varepsilon_{0B} \quad (5.18)$$

where *Y*, *X*, and ε are outcome, conditioning variables, and error term, respectively. But they have the subscripts "0" and "B" that means "no program" and "before the program", respectively. Before the program, all people are in status of no program, and the observed outcome is the outcome in the absence of the program.

After the program, the denotation of the potential outcomes is similar to the case of single cross-section data, but has an additional subscript "A" that means "after the program":

$$Y_{0A} = \alpha_{0A} + X_A\beta_{0A} + \varepsilon_{0A} \quad (5.19)$$

$$Y_{1A} = \alpha_{1A} + X_A\beta_{1A} + \varepsilon_{1A} \quad (5.20)$$

Then, the conditional parameters of interest are expressed as follows:

$$ATE_{(X)} = (\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A}) + E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A) \quad (5.21)$$

$$ATT_{(X)} = (\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A}) + E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A, D = 1) \quad (5.22)$$

The key assumption in the first-difference method is that the error term includes a time-invariant component and any correlation between *D* and the error is included in this component. The time-invariant component can be called the fixed and unobserved effect.

Assumption 5.4: Error terms in the potential outcome equations are decomposed to components with the following properties:

$$\varepsilon_{0B} = \pi + \eta_{0B}, \qquad \varepsilon_{0A} = \pi + \eta_{0A}, \ \varepsilon_{1A} = \pi + \eta_{1A},$$

where:

$$\eta_{0B}, \eta_{0A}, \eta_{1A} \perp D \mid X_B, X_A \qquad \text{(A.5.4)}^{13}$$

In addition, to identify $ATE_{(X)}$ and $ATT_{(X)}$, we need assumptions on exogeneity of X, i.e., an assumption similar to (A.2.1):

Assumption 5.5: $E(\varepsilon_{0B} \mid X_B, X_A) = E(\varepsilon_{0A} \mid X_B, X_A) = E(\varepsilon_{1A} \mid X_B, X_A) = 0$ (A.5.5)

Proposition 5.4: Under assumptions (A.5.4) and (A.5.5), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified and can be estimated by OLS regression.

Proof:

Firstly, under assumption (A.5.4) and (A.5.5), $ATE_{(X)}$ and $ATT_{(X)}$ are identified and the same, since:

$$E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A) = 0,$$

$$\begin{aligned} E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_B, X_A, D = 1) &= E(\eta_{1A} - \eta_{0A} \mid X_B, X_A, D = 1) \\ &= E(\eta_{1A} - \eta_{0A} \mid X_B, X_A) \\ &= E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_B, X_A) \\ &= 0, \end{aligned}$$

As a result, $E(\varepsilon_{1A} - \varepsilon_{0A} \mid X_A, D = 1) = 0$.

The estimator of $ATE_{(X)}$ and $ATT_{(X)}$ is the coefficient of D in the following equation:

$$Y_A = \alpha_{0A} + X_A\beta_{0A} + D\left[(\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A})\right] + \left[D(\varepsilon_{1A} - \varepsilon_{0A}) + \varepsilon_{0A}\right] \quad (5.23)$$

To estimate $(\alpha_{1A} - \alpha_{0A})$ and $(\beta_{1A} - \beta_{0A})$, subtract (5.19) from (5.25) to obtain:

$$\begin{aligned} Y_A - Y_{0B} = (\alpha_{0A} - \alpha_{0B}) + (X_A\beta_{0A} - X_B\beta_{0B}) + D\left[(\alpha_{1A} - \alpha_{0A}) + X_A(\beta_{1A} - \beta_{0A})\right] \\ + \left[D(\varepsilon_{1A} - \varepsilon_{0A}) + (\varepsilon_{0A} - \varepsilon_{0B})\right], \end{aligned} \quad (5.24)$$

in which the error term has the traditional property due to the (A.5.4) and (A.5.5):

$$\begin{aligned} & E\left\{\left[D(\varepsilon_{1A} - \varepsilon_{0A}) + (\varepsilon_{0A} - \varepsilon_{0B})\right] X_B, X_A, D\right\} \\ &= DE(\varepsilon_{1A} - \varepsilon_{0A} \mid X_{BA}, D) + E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}, D) \\ &= DE(\eta_{1A} - \eta_{0A} \mid X_{BA}, D) + E(\eta_{0A} - \eta_{0B} \mid X_{BA}, D) \\ &= DE(\eta_{1A} - \eta_{0A} \mid X_{BA}) + E(\eta_{0A} - \eta_{0B} \mid X_{BA}) \\ &= 0 \end{aligned} \quad (5.25)$$

Thus, we can estimate all coefficients in (5.24) without bias by running regression of the difference in observed outcome before and after the program on X_B and X_A, and the program selection variable D. Then, the estimates of these

[13] In some econometrics text, $\eta_{0B}, \eta_{0A}, \eta_{1A} \perp D \mid X_B, X_A$ is called strict exogeneity condition.

coefficients will be used to estimate the conditional and unconditional parameters of the program impact.

Difference-in-difference with matching method

The method of difference-in-difference with matching can be regarded a non-parametric version of the first-difference method. It allows the program selection to be based on unobservable variables in sense that it does not require the conditional independence assumption (A.4.1). However, it requires the bias be time-invariant. Compared with the first-difference method, it has an advantage that it does require the assumption on exogeneity of *X* to identify the program impact parameters and it can be used without panel data.

Proposition 5.4: Under assumptions (A.5.4), $ATE_{(X)}$, $ATT_{(X)}$, ATE and ATT are identified and can be estimated non-parametrically by the matching method.

Proof:

From (A.5.4), we get:

$$\begin{aligned} E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_B, X_A, D) &= E(\eta_{0A} - \eta_{0B} \mid X_B, X_A, D) \\ &= E(\eta_{0A} - \eta_{0B} \mid X_{BA}) \\ &= E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}), \end{aligned} \tag{5.26}$$

where X_{BA} denote all X_B and X_A. Thus, $E(\varepsilon_{0A} - \varepsilon_{0B})$ is independent of *D* given X_B and X_A before and after the program. As a result:

$$E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}, D = 0) = E(\varepsilon_{0A} - \varepsilon_{0B} \mid X_{BA}, D = 1), \tag{5.27}$$

and we have:

$$E(Y_{0A} \mid X_{BA}, D = 0) - E(Y_{0B} \mid X_{BA}, D = 0) = E(Y_{0A} \mid X_{BA}, D = 1) - E(Y_{0B} \mid X_{BA}, D = 1) \tag{5.28}$$

Recall that $ATT_{(X)}$ is equal to:

$$ATT_{(X_B, X_A)} = E(Y_{1A} \mid X_{BA}, D = 1) - E(Y_{0A} \mid X_{BA}, D = 1). \tag{5.29}$$

Insert (5.28) into (5.29) to obtain:

$$\begin{aligned} ATT_{(X_B, X_A)} &= E(Y_{1A} \mid X_{BA}, D = 1) - E(Y_{0A} \mid X_{BA}, D = 1) - \left[E(Y_{0A} \mid X_{BA}, D = 0) - E(Y_{0B} \mid X_{BA}, D = 0)\right] \\ &\quad + \left[E(Y_{0A} \mid X_{BA}, D = 1) - E(Y_{0B} \mid X_{BA}, D = 1)\right] \\ &= \left[E(Y_{1A} \mid X_{BA}, D = 1) - E(Y_{0A} \mid X_{BA}, D = 0)\right] - \left[E(Y_{0B} \mid X_{BA}, D = 1) - E(Y_{0B} \mid X_{BA}, D = 0)\right] \end{aligned}$$

Similarly, we can identify the conditional average effect of non-treatment on the non-treated (ANTT):

$$\begin{aligned} ANTT_{(X_B, X_A)} &= E(Y_{1A} \mid X_{BA}, D = 0) - E(Y_{0A} \mid X_{BA}, D = 0) - \left[E(Y_{1A} \mid X_{BA}, D = 0) - E(Y_{0B} \mid X_{BA}, D = 0)\right] \\ &\quad + \left[E(Y_{1A} \mid X_{BA}, D = 1) - E(Y_{0B} \mid X_{BA}, D = 1)\right] \\ &= \left[E(Y_{1A} \mid X_{BA}, D = 1) - E(Y_{0A} \mid X_{BA}, D = 0)\right] - \left[E(Y_{0B} \mid X_{BA}, D = 1) - E(Y_{0B} \mid X_{BA}, D = 0)\right], \end{aligned}$$

which is the same as $ATT_{(X)}$. As a result, $ATE_{(X)}$ is identified, and it is equal to $ATT_{(X)}$.

Participants are matched with non-participants based on their conditioning variables before and after the program, X_B and X_A. The above matching method requires panel data. If only independently pooled cross section data are available, the matching will be performed in a slightly different way. Firstly participants are matched to non-participants based on X_B to estimate the difference in their outcome before the program. Secondly, after the program participants are matched to non-participants again but based on X_A to estimate the difference in their outcome. Then, the estimate of the program impact $ATT_{(X)}$ is equal to the difference in the estimates before and after the program. That is why this method is also called double-matching.

The main advantage of the panel data methods is that it allows for the selection of the program based on time-invariant unobservable variables. However, the methods have two disadvantages. The first is the requirement of the data set. Panel data that are collected before and after the program are not always available. The second is that the methods require on a rather strong assumption that unobservable variables that affect the program selection are unchanged over time.

6. Conclusions

There is a growing interest in impact evaluation of programs and policies from not only academic researchers but also policy makers. Estimation of the impact of a program is often challenging because of self-selection bias. Participants in the program are not randomly selected. They are selected in the program based on their decisions and program administrators' decisions. Different methods in impact evaluation rely on different assumptions on the relation between the outcome process and the program selection process to construct the counterfactual so that the program impacts are identified. Understanding these identification assumptions of impact evaluation methods and the selection process of programs helps researchers and evaluation practitioners select the relevant methods to measure the impact of programs.

The paper discusses alternative methods in terms of identification assumptions and estimation strategies in contexts of the two potential outcome equations and program selection equations with the allowance for heterogeneous program impacts. Ideally a program is randomly assigned to beneficiaries and the impact of the program is simply measured by the difference between outcomes of beneficiaries and outcome of non-beneficiaries. Although randomized studies are costly and require strict monitoring, the number of studies using randomized control trails has been increasing in the recent years because of its high internal validity in impact evaluation.

Most development programs and policies are not randomized. However, if the selection process of participants into a program is fully observed, we can estimate the program impact by running regression of outcomes on the variable of program participation and other control variables which affect the program participation and outcomes. The program impact can also be estimated using matching methods.

In reality, participants are often self-selected in programs. They decide to join the programs based on their own criteria and these criteria cannot be measured by the impact evaluation practitioners. In these cases, instrumental-variable regression, fixed-effects regression, and difference-in-differences estimators are widely used methods to measure the program impacts.

Conduction of rigorous impact evaluation is very costly. If we are interested in the causal effect of a program, well-designed impact evaluations should be carried out from the beginning of the program. Impact evaluation should be understood as a continuous process during the program implementation. Control and treatment

groups should be designated before the program start. They need to be tracked so that the selection process and problems of attrition and substitution can be fully observed. Finally, baseline and post-project surveys need to be conducted using the same survey instruments to ensure the comparison.

References

Abadie, A. & Imbens, G.W. (2002). Simple and bias-corrected matching estimators for average treatment effects. *NBER Technical Working*, Paper No. 283. doi. 10.3386/t0283

Abhijit, V., Banerjee, A., & Duflo, E. (2008). The experimental approach to development economics. *NBER Working Paper*, 14467. doi. 10.3386/w14467

Angrist, D.J., & Pischke J.S. (2009), *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Asian Development Bank, (2011), *A Review of Recent Developments in Impact Evaluation*, Asian Development Bank, Philippines.

Banerjee, A., & Duflo, E. (2011), *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*, Public Affairs Books, New York.

Blundell, R., & Costa-Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 4(3), 565-640. doi. 10.3368/jhr.44.3.565

Cochran, W.G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, 24(2), 295-313. doi. 10.2307/2528036

Cochran, W.G., & Chambers, S.P. (1965). The Planning of Observational Studies of Human Population. *Journal of the Royal Statistical Society*, 128(2), 234-266.doi. 10.1016/0021-9681(79)90034-1

Dawid, A.P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society*. 41(1), 1-31.

Dehejia, R.H. & Wahba, S. (1998). Propensity score matching methods for non-experimental causal studies. *NBER Working Paper*, No. 6829. doi. 10.3386/w6829

Duflo, E., Glennerster, R., & Kremer, M. (2008). Using Randomization in Development Economics Research: A Toolkit, Chapter 61. *Handbook of Development Economics*, Volume 4.

Fan, J. (1992). Local linear regression smoothers and their minimax efficiencies. *The Annals of Statistics*, 21(1), 196-216.doi. 10.1214/aos/1176349022

Greene, W.H. (2003). *Econometric Analysis.* Firth Edition, Prentice Hall Press.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2), 315-331. doi. 10.2307/2998560

Hahn, J., Todd, P., & van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1), 201-209. doi. 10.1111/1468-0262.00183

Heckman, J. (1978). Dummy endogenous variables in a simultaneous equation system. *Econometrica*, 46(4), 931-959. doi. 10.2307/1909757

Heckman, J., & Vytlacil, E.J. (1999). Local instrumental variables and latent variable models for identifying and bounding treatment effects. *Proceeding of National Academy of Science*, 96(8), 4730-4734. doi. 10.1073/pnas.96.8.4730

Heckman, J., Ichimura, H., & Todd, P.E. (1997b). Matching as an econometric evaluation estimators: evidence from evaluating a job training programme. *Review of Economic Studies*, 64(4), 605-654. doi. 10.2307/2971733

Heckman, J., Ichimura, H., Smith, J.A., & Todd, P.E. (1998a). Characterizing selection bias using experimental dada. *Econometrica*, 66(5), 1017-1098. doi. 10.2307/2999630

Heckman, J., Smith, J., & Clements, N. (1997a). Making the most out of program evaluations and social experiments: Accounting for heterogeneity in program impacts. *Review of Economic Studies*, 64(4), 487-535. doi. 10.2307/2971729

Heckman, J., Lochner, L., & Taber, C. (1998b). General equilibrium treatment effects: A study of tuition policy. *American Economic Review*. 88(2), 381-386.

Heckman, J., Hohmann, N., Khoo, M., & Smith, J. (2000). Substitution and dropout bias in social experiments: Evidence from an influential social experiment. *The Quarterly Journal of Economics*, 115(2), 651-694. doi. 10.1162/003355300554764

Heckman, J., Lalonde, R., & Smith, J. (1999). The economics and econometrics of active labor market programs. *Handbook of Labor Economics, Volume 3*, Ashenfelter, A. and *D*. Card, eds., Amsterdam: Elsevier Science.

Hirano, K., Imbens, G.W., & Ridder, G. (2002). Efficient estimation of average treatment effects using the estimated propensity score. *NBER Working Paper*, No.251. doi. 10.3386/t0251

Imbens, G.W., & Angrist, J.D. (1994). Identification and estimation of local average treatment effect. *Econometrica*, 62(2), 467-475. doi. 10.2307/2951620

Imbens, G.W., & Wooldridge, J.M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1), 5-86. doi. 10.1257/jel.47.1.5

Imbens, G.W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, 142(2), 615-635.doi.10.1016/j.jeconom.2007.05.001

Lee, D.S., & Lemieux, D.S. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, 48(2), 281-355.doi. 10.1257/jel.48.2.281

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proc. Nat. Inst. Sci. Ind.* 12 (1936), 49-55.

Moffitt, R. (1991). Program evaluation with nonexperimental data. *Evaluation Review*, 15(3), 291-314. doi. 10.1177/0193841X9101500301

Powell, J. (1994). Estimation of semiparametric models. *in*: R. Engle and *D*. McFadden, eds., *Handbook of Econometrics*, vol. 4 (North-Holland, Amsterdam, Netherlands), 2443-2521.

Rosenbaum, P., & Rubin, R. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41-55. doi. 10.1093/biomet/70.1.41

Rosenbaum, P., & Rubin, R. (1984). Reducing bias in observation studies using subclassification on the propensity score. *Journal of Statistical Association*, 79(387), 516-523.

Rosenbaum, P., & Rubin, R. (1985a). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician*, 39(1), 33-38.

Rosenbaum, P., & Rubin, R. (1985b). The bias due to incomplete matching. *Biometrics*, 41(1), 103-116. doi. 10.2307/2530647

Rubin, D. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66(5), 688-701. doi. 10.1037/h0037350

Rubin, D. (1977). Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics*, 2(1), 1-26. doi. 10.3102/10769986002001001

Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1), 34-58. doi.10.1214/aos/1176344064

Rubin, D. (1979). Using multivariate sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74, 318–328. doi. 10.1080/01621459.1979.10482513

Rubin, D. (1980). Bias reduction using Mahalanobis-Metric matching. *Biometrics*, 36(2), 293–298. doi. 10.2307/2529981

Smith, J. & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators?. *Journal of Econometrics*, 125(1–2), 305–353. doi. 10.1016/j.jeconom.2004.04.011

Staiger, D., & Jame, H.S. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–86. doi. 10.2307/2171753

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, 43(4), 1249-87. doi. 10.1111/1468-2354.t01-1-00055

White, H, (2006). *Impact Evaluation: the Experience of the World Bank's Independent Evaluation Group*. Washington, DC: World Bank.

White, H, (2009). *Some Reflections on Current Debates in Impact Evaluation*. Working Paper No. 1, International Initiative for Impact Evaluation.

Wooldridge, J.M. (2001). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts London, England.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by-nc/4.0).